REMARKS

The claims have been amended to clarify the invention. Claim 1 has been amended to delete "fragment" language in 1(c) and 1(d). Claim 5 has been amended to recite "An isolated host cell ...". Claim 10 has been amended to replace the term "differentially expressed" with "increased expression ...in the sample", and the term "standard" has been further defined as "standard of normal tissue". Support for the amendments to claim 10 are found in the specification at, for example, page 23- 29, and Figure 3B (increased expression in clear cell sarcoma), and at page 18, lines 25-32, which describes the methods of determining "standard values" from both normal and diseased tissues. Claim 12 has been amended to delete the terms "peptide nucleic acids" and "regulatory molecules". No new matter is added by any of these amendments, and entry of the amendments is therefore requested.

The Examiner stated that the rejection of claim 11 under 35 U.S.C. § 112, second paragraph, as set forth in the previous Office Action mailed December 31, 2002 has been withdrawn in view of Applicants' amendment to the claim.

Objection to the Specification

The Examiner has objected to the specification because the amendment filed April 2, 2003 introduces new matter into the disclosure. Specifically, the Examiner stated that the added material which is not supported by the original disclosure is the four ankyrin domains that differ from the originally disclosed domains in the paragraph beginning at the bottom of page 10, line 30. The Examiner stated that since the original ankyrin domains clearly differ from the amended ankyrin domains in their amino acid sequences, there is no support for in the original disclosure for the amended ankyrin domains, and applicants were not in possession of the amended ankyrin domains. Applicant is required to cancel the new mater in the reply to this Office Action.

Applicants Response

Applicants have argued in support of the above referenced amendment that the correct amino acid numbering for the residues recited in these domains is found in the Sequence Listing for SEQ ID NO:1 (Ankrd2V) as well as in Figures 2A-2D. Therefore, the renumbering of these residues in the specification merely conforms one part of the specification with another and is not new matter. Withdrawal of the objection is therefore requested.

35 U.S.C. § 101, Rejection of Claim 5

The Examiner stated that Claim 5 recites a host cell comprising a vector. Thus, the claim reads on a transgenic human, which is non-statutory subject matter. It is recommended that "a host cell" be replaced by "an isolated host cell" to overcome the rejection. Claim 5 has been so amended. Withdrawal of the rejection is therefore requested.

35 U.S.C.§ 112, First Paragraph, Rejection of Claim 10

The Examiner has maintained the rejection of claim 10 under 35 U.S.C. § 112, first paragraph (Enablement), for the reasons of record in the previous Office Action. The Examiner stated that since the specification clearly describes the upregulation or increased expression of Ankrd2V in clear cell sarcoma of the skeletal relative to normal muscle tissue, one skilled in the art would clearly interpret the use of the term "differentially expressed" as recited in the claim to refer only to the increased or upregulation of the gene in the diseased state. The Examiner stated that Applicants' arguments have been considered, but are not deemed persuasive. The Examiner stated that the disclosure only enables a portion of the term "an increased, upregulated". While it is true that the proper interpretation of the term should be in light of the knowledge of one skilled in the art and the disclosure of the specification, the disclosure simply fails to enable and artisan to practice the invention commensurate with the scope of the claim.

The Examiner also stated that applicants arguments regarding the definition of "a standard" and its clear meaning in the art has also been considered but is not deemed persuasive for the reasons set forth in the previous Office Action. The Examiner stated that a standard indicates a specific numerical value that is obtained by comparing to "normal" and applicants fail to disclose that particular value.

Applicants Response

Applicants believe that the term "differential expression " as recited in the claim together with the data related to differential expression, in particular, upregulation in clear cell sarcoma, as presented in Figure 3B would enable one skilled in the art to practice the invention commensurate with the scope of the claim. However, in the interest of expediting prosecution and the allowance of the claim, the term "differential expression" has been replaced with "increased expression".

With respect to the Examiners' insistence that "a standard" must carry a specific

numerical value, applicants strongly disagree. A specific numerical value for a given standard in any experiment can only be determined in each individual experiment. The specification clearly teaches that a "standard" for this method corresponds to a value obtained from "normal" (nondiseased) tissue. The Examiner appears to acknowledge this in the statement quoted above. In addition, the specification clearly teaches how to determine a "standard value" for any given experiment for normal or diseased tissue at page 18, lines 27-31. Therefore to further clarify this in the claim, the term a "standard" has been replaced with the phrase "a standard of normal tissue".

With these amendments, applicants submit that claim 10 is fully enabled commensurate with the scope of the claim and withdrawal of the rejection is therefore requested.

35 U.S.C. § 112, First Paragraph, Rejection to Claims 1-12

The Examiner has rejected claims 1-12 under 35 U.S.C. § 112, first paragraph (Enablement), because the specification does not reasonably provide enablement for one skilled in the art to use the invention commensurate in scope with the claims. Claim 1(b) recites an isolated cDNA comprising a naturally occurring variant of the amino acid sequence of SEQ ID NO:1. Claim 1(c) recites an isolated cDNA comprising a biologically active fragment of SEQ ID NO:1. Claim 2(b) recites a fragment of SEQ ID NO:2 selected from SEQ ID NOs:3-6, and claim 2(c) recites a variant of SEQ ID NO:2 selected from SEQ ID NOs:7-10. Therefore, the Examiner stated, the claims encompass a genus comprising an enormous number of nucleic acids which vary greatly both in length and nucleic acid composition. However, other than the cDNA comprising the nucleic acid sequence of SEQ ID NO:2 that encodes SEQ ID NO:1, the disclosure fails to provide sufficient guidance and information regarding the structural and functional requirements commensurate in scope with what is encompassed by the instant claims. The disclosure has not shown which portions of SEQ ID NO:2 are critical to the activity of the Ankrd2V of SEQ ID NO:1, or what modifications (e.g., substitutions, deletions or additions) one can make to SEQ ID NO:2 will result in protein mutants with the same functions as the protein of SEQ ID NO:2.

Applicants Response

Claim 1(c) has been amended to delete fragment language, therefore the rejection is moot with respect to these elements of the claim. With respect to the claimed variants encoding SEQ

ID NO:1, and fragments and variants of SEQ ID NO:2, the use of all polynucleotides of the invention are fully enabled to one of skill in the art throughout the specification.

For example, the polynucleotides of the invention are enabled as hybridization probes for the diagnosis of disease conditions associated with differential expression of Ankrd2V, in particular, clear cell sarcoma (specification at p. 4, lines 6-11), as well as for identifying naturally occurring molecules encoding Ankrd2V, allelic variants, or related molecules (p. 13, lines 22-26); in arrays to monitor large number of genes simultaneously and to identify genetic variants and mutations (p. 14, lines 16-24); and for chromosomal mapping (p. 14, lines 25-29). Furthermore, all of the mammalian cDNA variants are enabled for producing transgenic cell lines or organisms which model human disorders and upon which potential therapeutic treatments may be tested, in particular for human clear cell sarcoma (p. 11, lines 30-31, and p. 12, lines 6-8). None of the described uses of the polynucleotides requires a functional association of an encoded polypeptide.

The Examiner further stated that claim 2(b) recites a fragment of SEQ ID NO:2, however, that a sequence alignment shows that SEQ ID NO:3 is not exactly a fragment of SEQ ID NO:2. Furthermore, the Examiner stated, claim 2(c) recites a variant of SEQ ID NO:2 selected from SEQ ID NO:7-10 and sequence analysis shows that they share some similarity with only a small portion of SEQ ID NO:2. The instant disclosure asserts these cDNAs are particularly useful for producing transgenic cell lines or organisms which model human disorders upon which potential therapeutic(s) for such disorders may be tested, however, no specific disorders are disclosed (emphasis added).

Applicants Response

Given that the Examiner's alignment indicates the SEQ ID NO:3 and SEQ ID NO:2 may differ by the alignment of two base pairs out of 420, the Examiner presents no evidence whatsoever why one skill in the art could not use the claimed sequence as a probe for SEQ ID NO:2 without undue experimentation, or for any of the other uses presented above, particularly considering that the exact sequence of SEQ ID NO:3 is disclosed.

Likewise, while the mammalian variants obviously represent variants of a portion of SEQ ID NO:2 (see column 5 of the Table at p. 11), the Examiner presents no evidence as to why one skilled in the art would require undue experimentation to practice the use of said variants in, for

example, producing transgenic cell lines or organisms for modeling human diseases associated with Ankrd2V expression, in particular, human clear cell sarcoma. With regard to the Examiner's statement that no specific disorders for this purpose are disclosed, applicants submit that it is stated throughout the specification that human clear cell sarcoma is associated with increased Ankrd2V expression, and therefore that one skilled in the art would clearly understand clear cell sarcoma to be a "disorder upon which potential therapeutics may be tested" with the instant polynucleotides. Furthermore, it is specifically recited in the specification at p. 12, lines 6-8 that; "The mammalian cDNAs may be used to produce transgenic cell lines or organisms which are model systems for human clear cell sarcoma and upon which the toxicity and efficacy of potential therapeutic treatments may be tested" (emphasis added).

The Examiner also stated that claim 6 recites a method for using a cDNA to produce a protein comprising culturing a host cells comprising a vector comprising the cDNA. However, the Examiner stated, it is well known in the art that only when an "expression vector" (emphasis added) comprising the cDNA is used, will a protein be produced. The Examiner suggested that replacing "A vector" with "An expression vector" would overcome this part of the rejection.

Applicants Response

It would be abundantly clear to one of skill in the art from the specification and common knowledge in the art that "A vector" refers specifically to "An expression vector", particularly considering the detailed description on protein expression at pp. 14-16 of the specification. However, applicants would consider amending the claim as the Examiner suggests pending the resolution of other outstanding rejections of the claims.

In summary, applicants submit the polynucleotides of the invention are fully enabled to one skilled in the art based on the arguments presented above. Further, they do not represent a "myriad of nucleic acid molecules that vary substantially in length but also in nucleotide composition" as the Examiner suggests, because they either represent specific nucleotide sequences presented in the Sequence Listing (SEQ ID nos"2-12), or are limited to polynucleotides encoding proteins at least 90% identical to SEQ ID NO:1. Withdrawal of the rejection of claims 1-12 under 35 U.S.C. § 112, first paragraph for lack of enablement is therefore requested.

35 U.S.C. § 112, First Paragraph, Rejection of Claims 1 and 3-12 (New Matter)

The Examiner has rejected claims 1 and 3-12 under 35 U.S.C. § 112, first paragraph, as failing to comply with the written description requirement. The claim(s) contains subject matter which is not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.

The Examiner stated that claim 1(b) recites a naturally occurring variant of the amino acid sequence of SEQ ID NO:1 having at least 90% identity to SEQ ID NO:1. There is not support for [this recitation] in the instant disclosure (p. 4, lines 22-23). The rejection further addresses fragments of SEQ ID NO:1 which, however, have been deleted from the claims.

Applicants Response

The description of "variants" referenced at p. 4, lines 22-23 describes their identification in terms of BLAST score (at least 100, more preferably 400) and as allelic variants having a "high percentage identity to the cDNAs and may differ by about three bases per hundred bases", descriptions which, to the skilled artisan, would convey a high (i.e., 85-90%) sequence identity to a given protein or cDNA. In addition, the specification further recites "a variant having at least 85% identity to the amino acid sequence of SEQ ID NO:1" at p. 4, lines 22-23 of the specification and, at p. 11, lines 6-7 and Figures 2A and 2D of the specification, discloses that SEQ ID NO:1 shares 88% identity with the ankyrin proteins, Ankrd2 (SEQ ID NO:11) and SMCP (SEQ ID NO:12). From these disclosures it would be readily apparent to one skilled in the art that applicants were in possession of a polypeptide within these parameters, i.e., at least 90% sequence identity to SEQ ID NO:1.

35 U.S.C. § 112, First Paragraph, Rejection of Claims 1-12 (Description)

The Examiner has rejected claims 1-12 under 35 U.S.C.§ 112, first paragraph, as containing subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor, at the time the application was filed, had possession of the claimed invention.

The Examiner stated that the description discloses a nucleic acid of SEQ ID NO:2 that encodes an ankyrin repeat domain 2 protein variant (Ankrd2V) of SEQ ID NO:1. However, claim 1 is drawn to an isolated cDNA comprising a nucleic acid sequence encoding a naturally

occurring variant of SEQ ID NO:1 having at least 90% identity to the amino acid sequence of SEQ ID NO:1. Claim 2 is drawn to an isolated cDNA comprising a fragment of SEQ ID NO:2 selected from SEQ ID NO:3-6 or a variant of SEQ ID NO:2 selected from SEQ ID NO:7-10. The claims do not require that the nucleic acids possess any particular biological activity (claims 1 and 2), nor any particular conserved structure, or disclosed distinguishing feature (claim1(b) and claim 2(c)). Thus, the Examiner stated, the claims are drawn to a genus of nucleic acids that is defined only by partial sequence identity.

The Examiner stated that factors to be considered if providing adequate written description and evidence of possession of a claimed genus include; complete or partial structure, physical and/or chemical properties, functional characteristics, structure/function correlation, methods of making the claimed product or any combination thereof. In this case, the Examiner stated, the only factor present in claim 1 and 2 is a partial structure of SEQ ID NO:1 or 2 and there is no functional limitation for the recited nucleic acids. The only factor present in claim 1(b) is a recitation of having at least 90% sequence identity with SEQ ID NO:1, and there is not even identification of any particular portion of the structure that must be conserved. Accordingly, the Examiner stated, in the absence of sufficient recitation of distinguishing identifying characteristics, the specification does not provide adequate written description of the claimed genus. The Examiner then cited various case law in support of his position, including *Vas-Cath Inc v. Mahurkar, Fiers v. Revel,* and *Amgen v. Chugai Pharmaceutical Co.* The rejection further addresses fragments of SEQ ID NO:1 which, however, have been deleted from the claims.

Applicants Response

As the Examiner as noted, the requirements necessary to fulfill the written description requirement of 35 U.S.C. 112, first paragraph, are well established by case law as well as by the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001. Applicants simply disagree with the Examiner's interpretation of these requirements. For example, *Vas-Cath Inc v. Mahurkar* states:

> . . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession *of the invention*. The invention is, for purposes of the "written description" inquiry,

whatever is now claimed. Vas-Cath, Inc. v. Mahurkar, 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

The Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001, provide that :

> An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, i.e., **complete or partial structure, other physical and/or chemical properties**, functional characteristics when coupled with a known or disclosed correlation between function and structure, **or some combination of such characteristics. What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail.** If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met (emphasis added).

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

SEQ ID NO:1 and SEQ ID NO:2 are specifically disclosed in the application (see, for example, page 3, lines 24-26). Variants of SEQ ID NO:1 are described, for example, at page 4, lines 22-23, where, in particular, a variant having at least 85% amino acid sequence similarity to SEQ ID NO:1 is described. Variants of SEQ ID NO:2 are specifically disclosed in terms of specific chemical structures at, for example, page 3, lines 27-28. Incyte clones in which the nucleic acids encoding the human Ankrd2V were first identified and libraries from which those clones were isolated are described, for example, at page 10, lines 18-22 of the Specification. Chemical and structural features of SEQ ID NO:1 are described, for example, on page 10, line 30 through page 11, line 14. Given SEQ ID NO:1, and the various chemical and structural features described for SEQ ID NO:1, one of ordinary skill in the art would recognize naturally-occurring variants of SEQ ID NO:1 having 90% sequence identity to SEQ ID NO:1. Accordingly, the Specification provides an adequate written description of the recited polypeptide sequences.

A. The Specification provides an adequate written description of the claimed "variants" of SEQ ID NO:1.

The Office Action has further asserted that the claims are not supported by an adequate written description because:

> The claims do not require that the nucleic acids possess any particular biological activity (claims 1 and 2), nor any particular conserved structure, or other disclosed distinguishing feature (claim 1(b) and claim 2(c)). Thus the claims are drawn to a genus of nucleic acids that is defined only by partial sequence identity.
>
> (page 9 of the Office Action)

Such a position is believed to present a misapplication of the law.

1. The present claims specifically define the claimed genus through the recitation of chemical structure

Court cases in which "DNA claims" have been at issue commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

> If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

> In claims to genetic material, however, a generic statement such as "vertebrate insulin cDNA" or "mammalian insulin cDNA," without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

> 1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

> A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. §112; *i.e.*, "an mRNA of a vertebrate, which mRNA encodes insulin" in *Lilly*, and "DNA which codes for a human fibroblast interferon-beta polypeptide" in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polynucleotides or polypeptides in terms of chemical structure, rather than on functional characteristics. For example, the "variant language" of independent claim 1 recites chemical structure to define the claimed genus:

> 1. An isolated cDNA ... comprising a nucleic acid sequence encoding a protein selected from the group consisting of:...b) a naturally-occurring variant of the amino acid sequence of SEQ ID NO:1 having at least 90% identity to the sequence of SEQ ID NO:1...

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. In the present case, there is no reliance merely on a description of functional characteristics of the polypeptides recited by the claims. In fact, there is no recitation of functional characteristics. Moreover, if such functional recitations were included, it would add to the structural characterization of the recited polypeptides. The polypeptides defined in the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry "on whatever is now claimed," the Office Action failed to

provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*

2. The present claims do not define a genus which is highly variant

Furthermore, the claims at issue do not describe a genus which could be characterized as highly variant. Available evidence illustrates that the claimed genus is of narrow scope.

In support of this assertion, the Examiner's attention is directed to the enclosed reference by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships," Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that ≥40% identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

The present application is directed, *inter alia*, to ankyrin repeat domain 2 proteins related to the amino acid sequence of SEQ ID NO:1. In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as ankyrin repeat domain 2 proteins and which have as little as 40% identity over at least 70 residues to SEQ ID NO:1. The "variant language" of the present claims recites, for example, polynucleotides encoding "a naturally-occurring amino acid sequence having at least 90% sequence identity to the sequence of SEQ ID NO:1" (note that SEQ ID NO:1 has 329 amino acid residues). This variation is far less than that of all potential ankyrin repeat domain 2 proteins related to SEQ ID NO:1, i.e., those ankyrin repeat domain 2 proteins having as little as 40% identity over at least 70 residues to SEQ ID NO:1.

3. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. §112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and

Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those case was based on the state of the art at essentially at the "dark ages" of recombinant DNA technology.

The present application has a priority date of April 26, 1999. Much has happened in the development of recombinant DNA technology in the 20 or more years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances one of skill in the art would recognize that, given the sequence information of SEQ ID NO:1, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed polynucleotide variants encoding SEQ ID NO:1at the time of filing of this application.

4. Summary

The Office Action failed to base its written description inquiry "on whatever is now claimed." Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polypeptides defined by claim 1 is adequately described, as evidenced by Brenner et al and consideration of the claims of the '740 patent involved in *Lilly*. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action. It is further noted that since the claimed variants of SEQ ID NO:2 are given in terms of their specific chemical structures (e.g., SEQ ID NO:7-10), these clearly meet the written description requirement.

Withdrawal of the rejection of claims 1-12 under 35 U.S.C.§ 112, first paragraph for lack of adequate written description is therefore requested.

35 U.S.C. § 112, Second Paragraph, Rejection of Claims 10 and 12

The Examiner has maintained the rejection of claims 10 and 12 under 35 U.S.C. § 112, second paragraph as set forth at page 4 of the previous Office Action (December 31, 2002). The Examiner stated that applicants arguments, at page 7 of the response filed 3/27/2003, have been considered but are not deemed persuasive because the instant disclosure fails to define unambiguously the term "a standard", an artisan would not know the metes and bounds of the term. A standard indicates a specific numerical value that is obtained by comparing to "normal". Applicants only provide guidance on how one may determine it, but fail to disclose that particular value.

Applicants Response

As previously stated in the response to the rejection of claim 10 under 35 U.S.C. § 112, first paragraph, above, a specific value for the term "a standard" could only be determined in an individual experiment. However, the specification describes how to determine such a value in any given experiment at, or example 18, lines 25-32 of the specification. The claim has, furthermore, been amended to replace the term "a standard" with the phrase "a standard of normal tissue" as the Examiner acknowledges the common definition of a standard to be, and which is described in the above referenced portion of the specification. One of skill in the art would clearly understand the metes and bounds of the term as recited in the amended claim.

The Examiner stated further that claim 12 is indefinite because it recites "peptide nucleic acids" and "regulatory molecules". Applicants arguments that the terms are well known in the art have been considered but are not deemed persuasive because the term "peptide nucleic acids" is not well known and the disclosure fails to define the term unambiguously. In addition the specification only provides exemplary examples of the term "regulatory molecules". Since neither the art nor the specification provides an unambiguous definition for the term, the claim is indefinite.

Applicants Response

Applicants maintain that the terms "peptide nucleic acids" and "regulatory molecules" are well known in the art and need not be defined to one of skill in the art. However, in the interests of expediting prosecution and the allowance of claims the terms "peptide nucleic acids" and "regulatory molecules" have been deleted from the claim.

With these amendments and arguments, applicants believe claims 10 and 12 are clear and definite and withdrawal of the rejection of these claims under 35 U.S.C. § 112, second paragraph is therefore requested.

35 U.S.C. § 112, Second Paragraph, Rejection of Claims 1-12

The Examiner has rejected claims 1-12 under 35 U.S.C. § 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

The Examiner stated that claim 1 is indefinite because it recites the term "biologically active fragment". Neither the art or the instant disclosure defines the term unambiguously and the term encompasses all biological activities, for example, from cell growth, differentiation, to immunological activities.

The Examiner stated that claim 1 is also indefinite for the recitation of the term "the complement thereof". It is well known in the art that cDNA encodes an amino acid sequence whereas its complement does not encode the same amino acid sequence.

The Examiner further stated that claim 2 is indefinite because SEQ ID NO:3 is not a fragment as the claim recites.

Applicants Response

Applicants disagree that the terms recited in the claims are indefinite and that one of skill in the art would not know their metes and bounds. Fragment language has been deleted from claim 1, therefore the rejection is moot with respect to the definition of "a biologically active fragment". With respect to the term "the complement thereof" of a cDNA encoding SEQ ID NO:1, it is further moot as to whether or not the claimed cDNA encodes the same amino acid sequence as that encoding SEQ ID NO:1,or any amino acid sequence at all. "[T]he complement" of a cDNA encoding SEQ ID NO:1 is unambiguously determined to one of skill in the art from the well known base pairing of double stranded DNA, i.e., A to T and G to C, and the degeneracy of the amino acid code. The claim is therefore clear and definite based on this common knowledge and the disclosure in the specification of SEQ ID NO:1 and of at least one cDNA encoding SEQ ID NO:1, i.e., SEQ ID NO:2.

With respect to the recitation of SEQ ID NO:3 as a fragment of SEQ ID NO:2 in claim 2, applicants disagree that one skilled in the art would not recognize SEQ ID NO:3 as an EST clone (Incyte Clone 972118R6) and an assemblage fragment from which the consensus sequence of SEQ ID NO:2 was derived, as stated at p. 10, lines18-22 of the specification, despite minor misalignments (i.e., 2 bases pairs out of 420). In any case, since the claim recites a specific nucleic acid sequence, i.e., SEQ ID NO:3, the metes and bounds of the claim element are unambiguously defined.

Applicants therefore submit that claims 1 and 2, as amended, are clear and definite and request withdrawal of the rejection of claims 1-12 under 35 U.S.C. § 112, second paragraph.

Claim Objections

The Examiner has objected to claim 1 because it (part a) recites "an amino acid sequence of SEQ ID NO:1". It appears, the Examiner stated, that "The" (instead of "an) should be used, and required appropriate correction.

Applicants disagree with this objection. Since the recitation of "an amino acid sequence of SEQ ID NO:1" is the first such recitation of the term in the claims, there is no antecedent basis for use of the definite article "the". Therefore, the indefinite article "an" is appropriate. Withdrawal of the objection is therefore requested.

CONCLUSION

In light of the above amendments and remarks, Applicants submit that the present application is fully in condition for allowance, and request that the Examiner withdraw the outstanding objections/rejections. Early notice to that effect is earnestly solicited.

If the Examiner contemplates other action, or if a telephone conference would expedite allowance of the claims, Applicants invite the Examiner to contact the undersigned at the number listed below.

Applicants believe that no fee is due with this communication. However, if the USPTO determines that a fee is due, the Commissioner is hereby authorized to charge Deposit Account No. **09-0108.**

Respectfully submitted,
INCYTE CORPORATION

Date: August 6, 2007

David G. Streeter, Ph.D.
Reg. No. 43,168
Direct Dial Telephone: (650) 845-5741

Customer No.: 27904
3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126
‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "lnscaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evoluti nary relationship though their sequence similarity may be low.

The recent growth f protein structure inf rmation combined with the comprehensive ev luti nary classificati n in the SCOP database (4, 5) have allowed us to overc me previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ~0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties −12/−1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

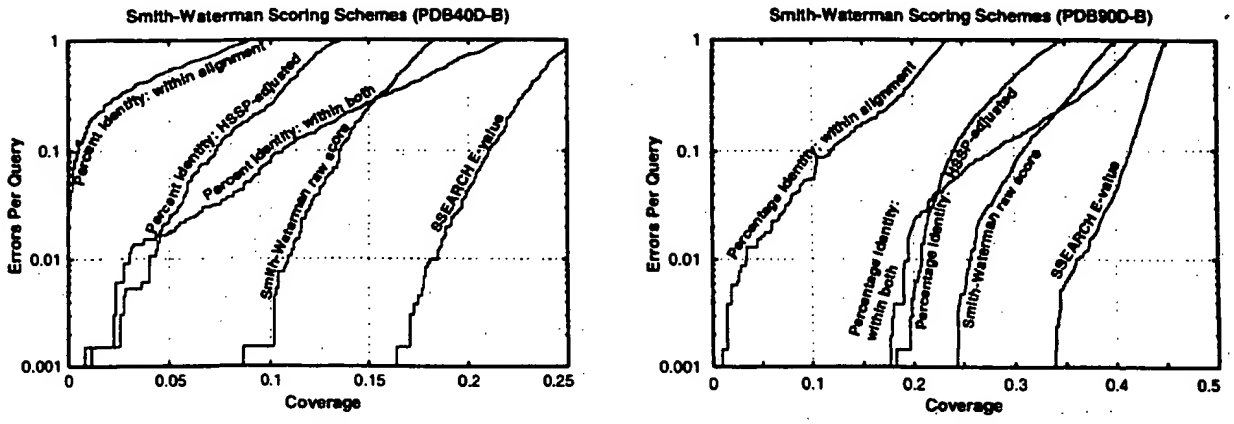

FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the x axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The y axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The y axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where l is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

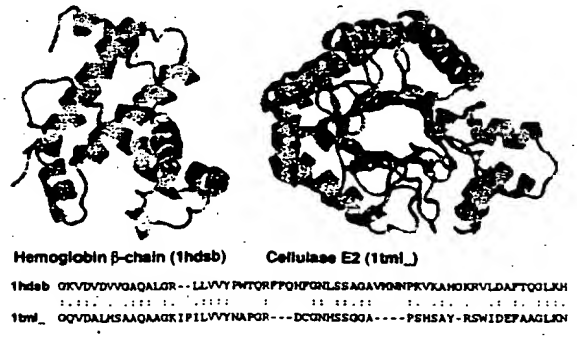

FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin β-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

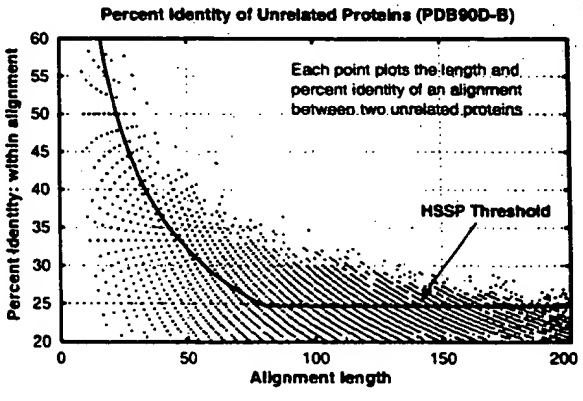

FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

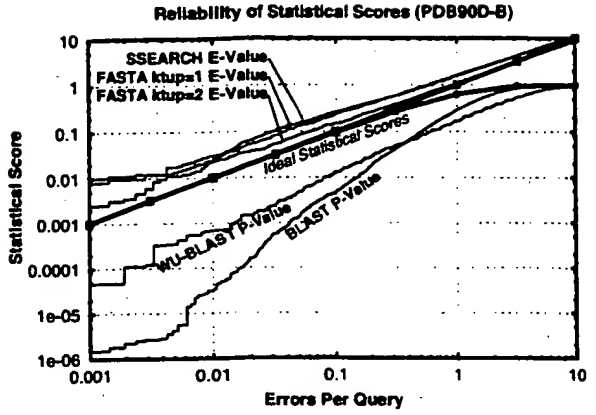

FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distributi n. These results are summarized in Fig. 1.

Sequence Identity. Th ugh it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

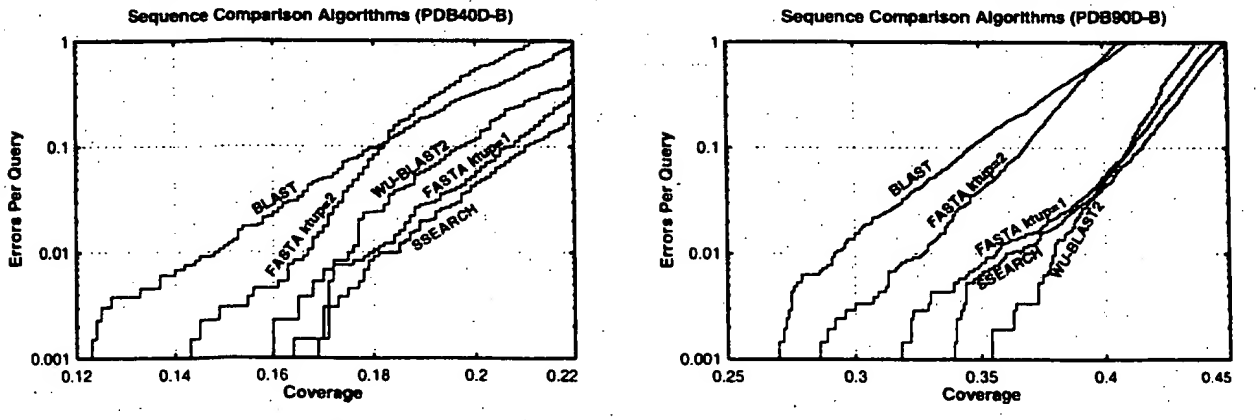

FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

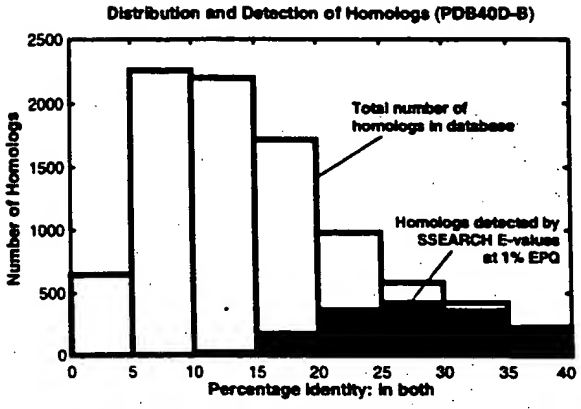

FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSSP-scaled	25.5	35% (HSSP + 9.8)	4.0
SSEARCH Smith–Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266**, 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266**, 635–643.
6. Pearson, W. R. (1991) *Genomics* **11**, 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4**, 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147**, 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266**, 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249**, 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17**, 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24**, 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24**, 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9**, 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233**, 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1**, 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1**, 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14**, 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266**, 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12**, 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266**, 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9**, 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13**, 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7**, 369–376.
32. Orengo, C., Michie, A., Jones S. Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5**, 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39**, 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20**, 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16**, 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4**, 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25**, 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131**, 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32**, 9906–9916
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20**, 374–376.